Asymptotic Minimax Regret for Data Compression, Gambling, and Prediction

Qun Xie and Andrew R. Barron, *Member, IEEE*

***Abstract*—For problems of data compression, gambling, and prediction of individual sequences $x_1, \cdots, x_n$ the following questions arise. Given a target family of probability mass functions $p(x_1, \cdots, x_n|\theta)$, how do we choose a probability mass function $q(x_1, \cdots, x_n)$ so that it approximately minimizes the maximum regret**

$$\max_{x_1,\cdots,x_n} (\log 1/q(x_1, \cdots, x_n) - \log 1/p(x_1, \cdots, x_n|\hat{\theta}))$$

and so that it achieves the best constant C in the asymptotics of the minimax regret, which is of the form $(d/2)\log(n/2\pi) + C + o(1)$, where d is the parameter dimension? Are there easily implementable strategies q that achieve those asymptotics? And how does the solution to the worst case sequence problem relate to the solution to the corresponding expectation version

$$\min_q \max_\theta E_\theta(\log 1/q(x_1, \cdots, x_n) - \log 1/p(x_1, \cdots, x_n|\theta))?$$

In the discrete memoryless case, with a given alphabet of size m, the Bayes procedure with the Dirichlet$(1/2, \cdots, 1/2)$ prior is asymptotically maximin. Simple modifications of it are shown to be asymptotically minimax. The best constant is

$$C_m = \log(\Gamma(1/2)^m/(\Gamma(m/2))$$

which agrees with the logarithm of the integral of the square root of the determinant of the Fisher information. Moreover, our asymptotically optimal strategies for the worst case problem are also asymptotically optimal for the expectation version.

Analogous conclusions are given for the case of prediction, gambling, and compression when, for each observation, one has access to side information from an alphabet of size k. In this setting the minimax regret is shown to be

$$\frac{k(m-1)}{2} \log \frac{n}{2\pi k} + kC_m + o(1).$$

***Index Terms*—Jeffreys' prior, minimax redundancy, minimax regret, universal coding, universal prediction.**

I. Introduction

E are interested in problems of data compression, gambling, and prediction of arbitrary sequences $x_1, x_2, \cdots, x_n$ of symbols from a finite alphabet $\mathcal{X}$. No probability distribution is assumed to govern the sequence. Nevertheless, probability mass functions arise operationally in the choice of data compression, gambling, or prediction strategies. Instead of a stochastic analysis of performance, our focus is the worst case behavior of the difference between the loss incurred and a target level of loss.

The following game-theoretic problem arises in the applications we discuss. We are to choose a probability mass function $q(x_1, \cdots, x_n)$ on $\mathcal{X}^n$ such that its conditionals $q(x_i|x_1, \cdots, x_{i-1})$ provide a strategy for coding, gambling, and prediction of a sequence x_i, $i = 1, 2, \cdots, n$. We desire large values of $q(x_1, \cdots, x_n)$ or equivalently small values of

$$\log 1/q(x_1, \cdots, x_n) = \sum_{i=1}^{n} \log 1/q(x_i|x_1, \cdots, x_{i-1})$$

relative to the target value achieved by a family of strategies. Specifically, let $\{p(x_1, \cdots, x_n|\theta), \theta \in \Theta\}$ be a family of probability mass functions on $\mathcal{X}^n$. One may think of it as a family of players, where the strategy used by player θ achieves value $\log 1/p(x_1, \cdots, x_n|\theta)$ for a sequence $x_1, \cdots, x_n$. Though we are not constrained to use any one of these strategies, we do wish to achieve for every $x_1, \cdots, x_n$ a value nearly as good as is achieved by the best of these players with hindsight. Thus the target level is $\log 1/p(x_1, \cdots, x_n|\hat{\theta})$ where $\hat{\theta} = \hat{\theta}(x_1, \cdots, x_n)$ achieves the maximum of $p(x_1, \cdots, x_n|\theta)$. The game-theoretic problem is this: choose q to minimize the maximum regret

$$\max_{x_1,\cdots,x_n} (\log 1/q(x_1, \cdots, x_n) - \log 1/p(x_1, \cdots, x_n|\hat{\theta})),$$

evaluate the minimax value of the regret, identify the minimax and maximin solutions, and determine computationally feasible approximate solutions. Building on past work by Shtarkov [30] and others, we accomplish these goals in an asymptotic framework including exact constants, in the case of the target family of all memoryless probability mass functions on a finite alphabet of size m.

The asymptotic minimax value takes the form $\frac{m-1}{2} \log \frac{n}{2\pi} + C_m + o(1)$, where C_m is a known constant. The choice of $q(x_1, \cdots, x_n)$ that is a mixture with respect to Jeffreys' prior (the Dirichlet$(1/2, \cdots, 1/2)$ in this case) is shown to be asymptotically maximin. A modification in which lower dimensional Dirichlet components are added near the faces of the probability simplex is shown to be asymptotically minimax. This strategy is relatively easy to implement using variants of Laplace's rule of succession. Moreover, unlike the exact minimax strategy, our strategies are also optimal for the corresponding expectation version of the problem studied in Xie and Barron [39].

The above game has interpretations in data compression, gambling, and prediction as we discuss in later sections. The choice of $q(x_1, \cdots, x_n)$ determines the code length

$$l(x_1, \cdots, x_n) = \log_2 1/q(x_1, \cdots, x_n)$$

Manuscript received January 16, 1997; revised June 4, 1998. This work was supported in part by NSF Under Grant ECS-9410760. The material in this paper was presented in part at the IEEE International Symposium on Information Theory, Ulm, Germany, 1997.

Q. Xie is with GE Capital, Stamford, CT 06927 USA.

A. R. Barron is with the Department Statistics, Yale University, New Haven, CT 06520 USA (e-mail: Andrew.Barron@yale.edu).

Communicated by N. Merhav, Associate Editor for Source Coding.

Publisher Item Identifier S 0018-9448(00)01361-4.

(rounded up to an integer) of a uniquely decodable binary code; it leads to a cumulative wealth

$$S_n(x_1,\cdots,x_n) = q(x_1,\cdots,x_n)O(x_1,\cdots,x_n)$$

after sequentially gambling according to proportions $q(x_{k+1}|x_1,\cdots,x_k)$ on outcome x_{k+1} with odds $O(x_{k+1}|x_1,\cdots,x_k)$ for $k=0,\cdots,n-1$; and, for prediction, a strategy based on $q(x_1,\cdots,x_n)$ incurs a cumulative logarithmic loss of

$$\log\left(1/q(x_1,\cdots,x_n)\right) = \sum_{k=0}^{n-1} \log 1/q(x_{k+1}|x_1,\cdots,x_k).$$

Likewise for each $p(x_1,\cdots,x_n|\theta)$ there is a code length $\log_2 1/p(x_1,\cdots,x_n|\theta)$, wealth $p(x_1,\cdots,x_n|\theta)O(x_1,|,x_n)$, and cumulative log loss $\sum_{i=0}^{n-1} \log 1/p(x_i|\theta)$.The target value corresponds to the maximum likelihood. The regret measures the difference in code lengths, the log wealth ratio, and the difference in total prediction loss between $q(x_1,\cdots,x_n)$ and the target level in the parametric family.

Recent literature has examined the regret for individual sequences in the context of coding, prediction, and gambling, in some cases building on past work on expected regret. Shtarkov [30] introduced and studied the minimax regret problem for universal data compression and gave asymptotic bounds of the form $(d/2)\log n+O(1)$ for discrete memoryless and Markov sources where d is the number of parameters. Extensions of that work to tree sources are in Willems, Shtarkov, and Tjalkens [38], see also [35] and [36]. Shtarkov *et al.* [31] identified the asymptotic constant in the regret for memoryless sources and addressed the issue of adaptation to an unknown alphabet. Rissanen [28] and Barron, Rissanen, and Yu [4] relate the stochastic complex criterion for model selection to Shtarkov's regret and show that the minimax regret takes the form $\frac{d}{2}\log n$ plus a constant identified under certain conditions (shown to be related to the constant that arises in the expectation version in Clarke and Barron [6]). Feder, Merhav, and Guttman [12], Haussler and Barron [17], Foster [14], Haussler, Kivinen, and Warmuth [18], Vovk [34], and Freund [15] studied prediction problems for individual sequences. Cover and Ordentlich ([7], [24]) presented a sequential investment algorithm and related it to universal data compression. Opper and Haussler [25] examine minimax regret for nonparametric problems.

Other related work considered expected regret. Davisson [9] systematically studied universal noiseless coding and the problem of minimax expected regret (redundancy). Davisson, McEliece, Pursley, and Wallace [11] as well as Krichevsky and Trofimov [22] identified the minimax redundancy to the first order. Other work giving bounds on expected redundancy includes Davisson and Leon-Garcia [10], Rissanen [26], [27], Clarke and Barron [5], [6], Suzuki [32], and Haussler and Opper [19].

Typically, the minimax expected regret with smooth target families with d parameters is of order $\frac{d}{2}\log n + C + o(1)$. The constant C and asymptotically minimax and maximin strategies for expected regret are identified in Clarke and Barron [6] (for the minimax value over any compact region internal to the parameter space) and in Xie and Barron [39] (for the minimax value over the whole finite-alphabet probability simplex). In these settings, [6] and [39] showed that the mixture with respect to the prior density proportional to $|I(\theta)|^{1/2}$ (Jeffreys' prior [20]) is asymptotically maximim.

In general, Bayes strategies for expected regret take the form of a mixture $q_W(x^n) = \int p(x_1,\cdots,x_n|\theta)W(d\theta)$, where W denotes a distribution on the parameter θ. In the expected regret setting (asymptotically) maximin procedures are based on a choice of prior W (or sequences of priors W_n) for which the average regret is (asymptotically) maximized [6]. Here we will seek choices of prior that yield asymptotically minimax values not just for the expected regret but also for the worst case pointwise regret. In addition to providing possibly natural probability assignment to the parameters the advantages of such a program are threefold. First, they afford ease of interpretation and computation (of predictions, gambles, and arithmetic codes) via the predictive distribution $q_W(x_i|x_1,\cdots,x_{i-1})$, not readily available for the exact minimax strategy of [30]. Secondly, the mixtures admit analysis of performance using information theory inequalities ([2], [3], [9]), and approximation by Laplace integration ([5], [6]).

Finally, achievement of an asymptotic regret not smaller than a specified value $\frac{d}{2}\log n + C + o(1)$ by a mixture strategy with a fixed prior W permits the conclusion that this is a pointwise lower bound for most sequences ([1], [4], [23], [35]). In particular, we find that for the class of memoryless sources, the Dirichlet $(1/2,\cdots,1/2)$ prior yields a procedure with regret possessing such a lower bound (Lemma 1), with what will be seen to be the minimax optimal value of C. Consequently, no sequence of strategies can produce regret much smaller than this for almost every data sequence (in a sense made precise in Section III below). These pointwise conclusions complement the result given below that the Dirichlet $(1/2,\cdots,1/2)$ mixture is asymptotically maximin.

One is tempted then to hope that the Dirichlet $(1/2,\cdots,1/2)$ mixture would also be asymptotically minimax for the simplex of memoryless sources. However, it is known that this mixture yields regret larger than the minimax level (by an asymptotically nonvanishing amount) for sequences that have relative frequencies near the boundary of the simplex (Lemma 3, in agreement with Suzuki [32] and Shtarkov [29]). Furthermore, Laplace approximation as in [6] suggests that this difficulty cannot be rectified by any fixed continuous prior. To overcome these boundary difficulties and to provide asymptotically minimax mixtures we use sequences of priors that give slightly greater attention near the boundaries to pull the regret down to the asymptotic minimax level. In doing so, the priors involve slight dependence on the target size n (or time horizon) of the class $\{p(x_1,\cdots,x_n|\theta),\theta\in\Theta\}$. Before specializing to a particular target family we state some general definitions and results in Section II. Among these are characterization of the minimax and maximin solution for each n and the conclusion that asymptotically maximin and asymptotically minimax procedures merge in relative entropy as $n\to\infty$. In Section III we examine the target class of memoryless sources over the whole probability simplex and identify an asymptotically minimax and maximin strategy based on a sequence of priors. Modifications to the Dirichlet $(1/2,\cdots,1/2)$ prior achieve these ob-

jectives and possess simple Laplace-type update rules. Proof of the asymptotic properties are given in Sections IV and V. Applications in gambling, prediction, and data compression are given in Sections VI, VII, and VIII.

Finally, in Section IX, we treat the problem of prediction (or gambling or coding) based on the class of more general models in which the observations may be predicted using a state variable (or side information) from an alphabet of size k. The asymptotic minimax regret is shown to equal

$$\overline{r}_n = \frac{k(m-1)}{2} \log \frac{n/k}{2\pi} + kC_m + o(1).$$

An asymptotically minimax procedure is to use a modified Dirichlet $(1/2, \cdots, 1/2)$ mixture separately for each state.

II. Preliminaries

Now we introduce some notation and preliminary results. Let a target family $\{p(x_1, \cdots, x_n|\theta), \theta \in \Theta\}$ be given. We occasionally abbreviate $(x_1, \cdots, x_n)$ to x^n and omit the subscript n from probability functions such as p_n, q_n, and $m_{w,n}$. Let the regret for using strategy $q_n(x^n)$ be defined by

$$r_n(q_n, x_1, \cdots, x_n) = \log \frac{p(x_1, \cdots, x_n|\hat{\theta})}{q_n(x_1, \cdots, x_n)}.$$

The minimax regret is

$$\overline{r}_n = \min_{q_n} \max_{x^n} r_n(q_n, x_1, \cdots, x_n).$$

A strategy q_n is said to be minimax if

$$\max_{x_1, \cdots, x_n} r_n(q_n, x_1, \cdots, x_n) = \overline{r}_n$$

and it is said to be an equalizer (constant regret) strategy if $r_n(q_n, x_1, \cdots, x_n) = \overline{r}_n$ for all $x_1, \cdots, x_n \in \mathcal{X}^n$. The maximin value of the regret is defined to be

$$\underline{r}_n = \max_{p_n} \min_{q_n} \sum_{x^n} p_n(x^n) r_n(q_n, x_1, \cdots, x_n)$$

where the maximum is over all distributions on $\mathcal{X}^n$. A strategy q_n is average case optimal with respect to a distribution p_n if it minimizes $\sum_{x^n} p_n(x^n) r_n(q_n, x^n)$ over choices of q_n. It is known from Shannon that the unique average case optimal strategy is $q_n(x^n) = p_n(x^n)$. A choice $q_n = p_n^*$ is said to be a maximin (or least favorable) strategy if

$$\sum_{x^n} r(p_n^*, x^n) p_n^*(x^n) = \underline{r}_n.$$

The following is basically due to Shtarkov [30] in the coding context.

Theorem 0: Let $c_n = \sum_{x^n} p(x^n|\hat{\theta})$ where $\hat{\theta} = \hat{\theta}(x^n)$ is the maximum-likelihood estimator. The minimax regret equals the maximin regret and equals

$$\overline{r}_n = \underline{r}_n = \log c_n.$$

Moreover, $q_n^*(x^n) = p(x^n|\hat{\theta})/c_n$ is the unique minimax strategy, it is an equalizer rule achieving regret

$$\log p(x^n|\hat{\theta})/q_n^*(x^n) = \log c_n$$

for all x^n, and it is the unique least favorable (maximin) distribution. The average regret for any other $p_n(x^n)$ equals

$$\sum_{x^n} p_n(x^n) \log(p(x^n|\hat{\theta})/p_n(x^n)) = \log c_n - D(p_n \| q_n^*).$$

We let $r_n = \overline{r}_n = \underline{r}_n = \log c_n$ denote the minimax = maximin value.

Proof of Theorem 0: Note that $\sum_{x^n} q_n^*(x^n) = 1$ and that $r_n(q_n^*, x^n) = \log c_n$ for all x^n, thus q_n^* is an equalizer rule. For any other $q(x^n)$ with $\Sigma_{x^n} q(x^n) = 1$, we must have $q(x^n) < q_n^*(x^n)$ for some x^n and hence

$$r_n(q_n, x^n) > r_n(q_n^*, x^n) = \log c_n$$

for that x^n. Thus q_n^* is minimax and $\overline{r}_n = \log c_n$. Now the last statement in the theorem holds by the definition of relative entropy and hence the maximin value

$$\begin{aligned} \underline{r}_n &= \max_{p_n} \sum r(p_n, x^n) p_n(x^n) \\ &= \max_{p_n} \sum p_n(x^n) \log \frac{p(x^n|\hat{\theta})}{p_n(x^n)} \\ &= \max_{p_n} (\log c_n - D(p_n \| q_n^*)) \end{aligned}$$

where $D(p_n \| q_n^*)$ is the relative entropy (Kullback-Leibler divergence). It is uniquely optimized at $p_n = q_n^*$, and therefore, $\underline{r}_n = \log c_n$. □

Thus the normalized maximum-likelihood $q_n^*(x^n) = p(x^n|\hat{\theta})/c_n$ is minimax. However, it is not easily implementable for online prediction or gambling which requires the conditionals, nor for arithmetic coding which also requires the marginals for $x_1, \cdots, x_k$, $k = 1, \cdots, n$. The marginals obtained by summing out $x_{k+1}, \cdots, x_n$ is not the same as $p(x^k|\hat{\theta}(x^k))/c_k$. See Shtarkov [30] for his comment on the difficulty of implementing q_n^* in the universal coding context.

In an asymptotic framework we can identify strategies that are nearly minimax and nearly maximin which overcome some of the deficiencies of normalized maximum likelihood. We say that a procedure $q_n(x^n)$ is asymptotically minimax if

$$\max_{x_1, \cdots, x_n} r_n(q_n, x_1, \cdots, x_n) = \overline{r}_n + o(1).$$

It is an asymptotically constant regret strategy if

$$r_n(q_n, x_1, \cdots, x_n) = \overline{r}_n + o(1)$$

for all x^n. A sequence $p_n(x^n)$ is asymptotically maximin if

$$\min_{q_n} \sum p_n(x^n) r_n(q_n, x^n) = \underline{r}_n + o(1).$$

It turns out that in general there is an information-theoretic merging of asymptotically maximin and minimax procedures in the sense stated in the following theorem.

Theorem 1: The Kullback–Leibler distance $D(p_n \| q_n^*)$ tends to zero as $n \to \infty$ for any asymptotically maximin p_n where q_n^* is the normalized maximum likelihood. Indeed, more generally

$$D(p_n \| q_n) \to 0$$

for any asymptotically maximin p_n and asymptotically minimax q_n.

Proof: From the last line of Theorem 0, $D(p_n\|q_n^*)$ measures how much below r_n is the average regret using p_n. Hence if p_n is asymptotically maximin then $D(p_n\|q_n^*) \to 0$. For any asymptotically maximin p_n and asymptotically minimax q_n we have

$$D(p_n\|q_n) = D(p_n\|q_n^*) + \sum p_n(x^n)\log(q_n^*(x^n)/q_n(x^n))$$

and $\max_{x^n} \log(q_n^*(x^n)/q_n(x^n))$ tends to zero by asymptotic minimaxity of q_n. So both terms in the above representation of $D(p_n\|q_n)$ tend to zero as $n \to \infty$. □

III. Main Result for Regret on the Simplex

Here we focus on the case that the target family is the class of all discrete memoryless sources on a given finite alphabet. In this case

$$p(x_1,\cdots,x_n|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$$

where $p(x = i|\theta) = \theta_i$, $i = 1,2,\cdots,m$, is the model of conditionally independent outcomes with $\theta = (\theta_1,\cdots,\theta_m)$ on the probability simplex

$$S_m = \left\{(\theta_1,\cdots,\theta_m) : \theta_i \geq 0 \text{ and } \sum_{i=1}^{m} \theta_i = 1\right\}.$$

The alphabet is taken to be $\mathcal{X} = \{1,2,\cdots,m\}$. Jeffreys' prior in this case is the Dirichlet $(1/2,\cdots,1/2)$ distribution. Earlier, Shtarkov [30] showed that the mixture with this prior achieves maximal regret that differs from the minimax regret asymptotically by not more than a constant.

Theorem 2: The minimax regret satisfies

$$r_n = \frac{d}{2}\log\frac{n}{2\pi} + C_m + o(1)$$

where $d = m - 1$ and $C_m = \log((\Gamma(1/2))^m/\Gamma(m/2))$. The choice $q(x^n) = m_J(x^n) = \int p(x^n|\theta)w_J(\theta)\,d\theta$ with $w_J(\theta)$ being the Dirichlet$_m(1/2,\cdots,1/2)$ prior is asymptotically maximin. It has asymptotically constant regret for sequences with relative frequency composition internal to the simplex. But it is not asymptotically minimax. The maximum regret on the boundary of the simplex is $r_n + (d/2)\log 2 + o(1)$, which is higher than the asymptotic minimax value. Finally, we give a modification of the Dirichlet $(1/2,\cdots,1/2)$ prior that provides a strategy of the form $\tilde{q}_n(x^n) = \int p(x^n|\theta)\,\tilde{W}_n(d\theta)$ that is both asymptotically minimax and maximin. Here $\tilde{W}_n = (1-\varepsilon_n)W_J + \varepsilon_n V$ is a mixture of Jeffreys' prior W_J on $(\theta_1,\cdots,\theta_m)$ and a small contribution from a prior $V = (1/m)\,\Sigma_{i=1}^m J_i$ with J_i on the lower dimension spaces

$$\left\{(\theta_1,\cdots,\theta_{i-1},1/n,\theta_{i+1},\cdots,\theta_m) : \sum_{i'\neq i}\theta_{i'} = 1 - 1/n\right\}$$

where $J_i = J_{i,n}$ makes

$$(\theta_1,\cdots,\theta_{i-1},\theta_{i+1},\cdots,\theta_m)/(1-\tfrac{1}{n})$$

have the Dirichlet$_{m-1}(1/2,\cdots,1/2)$ distribution and makes θ_i be fixed at $1/n$. Here $\varepsilon_n = n^{-1/8}$.

Remark 1: The above strategies $m_J(x^n)$ and $\tilde{q}_n(x^n)$ based on Jeffreys' prior and its modification, here shown to be asymptotically maximin and minimax for regret, are the same as shown to be asymptotically maximin and minimax for the expected regret in Xie and Barron [39] formulated there using expected regret defined by $E_\theta \log p(x^n|\theta)/q(x^n)$. Other satisfactory modifications of Jeffreys' prior are given in Section V.

In contrast, concerning the normalized maximum-likelihood strategy, though it is minimax for pointwise regret, it is not asymptotically minimax for expected regret as pointed out to us by Shtarkov. Indeed, the value $\max_\theta \; E_\theta \log p(x^n|\theta)/q_n^*(x^n)$ studied in Shtarkov [29] is asymptotically larger than the minimax expected regret identified in [39].

Remark 2: By asymptotic minimaxity the difference between the worst case regret of the strategy and the asymptotic value $(d/2)\log(n/2\pi) + C_m$ converges to zero with n (i.e., this difference is $o(1)$). We do not seek here to determine the optimal rate at which this difference converges to zero. Nevertheless, some bounds for it are given in Section V.

Remark 3: Jeffrey's mixture

$$m_J(x^n) = \int p(x^n|\theta)w_J(\theta)\,d\theta$$

can be expressed directly in terms of Gamma functions as

$$m_J(x^n) = D_m\left(T_{1,n}+\frac{1}{2},\cdots,T_{m,n}+\frac{1}{2}\right)\Big/ D_m\left(e\frac{1}{2},\cdots,\frac{1}{2}\right)$$

where $T_{i,n} = T_i(x^n)$ is the number of occurrences of the symbol i in $(x_1,\cdots,x_n)$, and

$$D_m(\lambda_1,\cdots,\lambda_m) = \prod_{i=1}^{m}\Gamma(\lambda_i)/\Gamma(\Sigma_{i=1}^m\lambda_i)$$

is the Dirichlet function. It can be more easily computed by the usual variant of Laplace's rule for conditionals. The conditionals $m_J(x_{k+1}|x_1,\cdots,x_k)$ are computed by

$$m_J(x_{k+1} = i|x_1,\cdots,x_k) = \frac{T_{i,k}+\frac{1}{2}}{k+\frac{m}{2}}$$

where $T_{i,k}$ is the number of occurrences of the symbol i in the sequence $(x_1,\cdots,x_k)$, and then

$$m_J(x_1,\cdots,x_n) = \prod_{k=0}^{n} m_J(x_{k+1}|x_1,\cdots,x_k).$$

Similarly, the asymptotically minimax (and maximin) strategy uses

$$\tilde{q}_n(x^n) = (1-\varepsilon_n)m_J(x^n) + \frac{\varepsilon_n}{m}\sum_{i=1}^{m} m_{i,n}(x^n)$$

where $m_J(x^n)$ is the Dirichlet mixture and

$$m_{i,n}(x^n) = \int p(x^n|\theta)J_{i,n}(d\theta)$$

is the mixture with the prior component $J_{i,n}$ in which $\theta_i = 1/n$ is fixed. Here $m_{i,n}(x^n)$ can be expressed directly as

$$\frac{D_{m-1}(T_1+\frac{1}{2},\cdots,T_{i-1}+\frac{1}{2},T_{i+1}+\frac{1}{2},\cdots,T_m+\frac{1}{2})}{D_{m-1}(\frac{1}{2},\cdots,\frac{1}{2})} \cdot\left(\frac{1}{n}\right)^{T_i}\left(1-\frac{1}{n}\right)^{n-T_i}.$$

This strategy $\tilde{q}_n$ can be computed by updating marginals according to

$$\tilde{q}_n(x^{k+1}) = \tilde{q}_n(x_{k+1}|x^k)\tilde{q}_n(x^k)$$

where the conditional probability is

$$\tilde{q}_{k,n}(x_{k+1}|x^k) = \frac{(1-\varepsilon_n)m_J(x^{k+1}) + \varepsilon_n \frac{1}{m}\sum_{i=1}^{m} m_{i,n}(x^{k+1})}{(1-\varepsilon_n)m_J(x^k) + \varepsilon_n \frac{1}{m}\sum_{i=1}^{m} m_{i,n}(x^k)} \quad (1)$$

and $m_J(x^k)$, $m_{i,n}(x^k)$ are updated according to

$$m_J(x^{k+1}) = m_J(x_{k+1}|x^k)m_J(x^k)$$

and

$$m_{i,n}(x^{k+1}) = m_{i,n}(x_{k+1}|x^k)m_{i,n}(x^k)$$

where

$$m_{i,n}(x_{k+1} = j|x_1,\cdots,x_k) = \begin{cases} \frac{T_{j,k}+1/2}{k-T_{i,k}+(m-1)/2}\left(1-\frac{1}{n}\right), & \text{for } j \neq i \\ \frac{1}{n}, & \text{for } j = i. \end{cases} \quad (2)$$

Therefore, simple recursive computations suffice. The total computation time is not more than the order of nm^2. Note, however, that our strategy requires knowledge of the time horizon n when evaluating the conditionals for x_{k+1} given $x_1,\cdots,x_k$ for $k = 0, 1, \cdots, n-1$ (also see Remark 9).

Remark 4: The answer $\frac{d}{2}\log\frac{n}{2\pi} + C_m$ is in agreement with the answer

$$\frac{d}{2}\log\frac{n}{2\pi} + \log\int_S \sqrt{|I(\theta)|}\,d\theta$$

that we would expect to hold more generally for smooth d-dimensional families with Fisher information $I(\theta)$, and parameter θ restricted to a set S, in accordance with Rissanen [28]. It also corresponds to the answer for expected regret from Clarke and Barron [6]. However, the present case of the family of all distributions on the simplex does not satisfy the conditions of [6] or [28].

Remark 5: Comparing r_n with the minimax value using expected loss in [39] and [6], $\frac{m-1}{2}\log\frac{n}{2\pi e} + \log\frac{\Gamma(1/2)^m}{\Gamma(m/2)} + o(1)$, we see that there is a difference of $\frac{m-1}{2}\log e$. The difference is due to the use in the expected loss formulation of a target value of $E_\theta \log 1/p(X^n|\theta)$ rather than $E_\theta \log 1/p(X^n|\hat{\theta})$, which differ by $E_\theta \log(p(X^n|\hat{\theta})/p(X^n|\theta)$, which, for θ internal to the simplex, is approximately one-half the expectation of a chi-square random variable with $m-1$ degrees of freedom. It may be surprising that in the present setting there is no difference asymptotically between the answers for minimax regret for individual sequences

$$r_n = \min_q \max_{x^n} \log p(x^n|\hat{\theta})/q(x^n)$$

and a minimax expected regret formulated here as

$$R_n = \min_q \max_\theta E_\theta \log p(x^n|\hat{\theta})/q(x^n).$$

In general, the minimax expected value R_n is less than the minimax pointwise regret r_n. To uncover situations when R_n and r_n agree asymptotically consider the maximin formulations of

$$R_n = \max_W \min_q \int W(d\theta)\sum_{x^n} p(x^n|\theta)\log p(x^n|\hat{\theta})/q(x^n)$$

and

$$r_n = \max_p \min_q \sum p(x^n)\log p(x^n|\hat{\theta})/q(x^n).$$

The difference is that in the former case the maximum is restricted to distributions of mixture type $\int W(d\theta)p(x^n|\theta)$. Asymptotically, R_n and r_n will agree if a sequence of mixture distribution is asymptotically least favorable (maximin) for the pointwise regret, as is the case in Theorem 2. Combining this conclusion with Remark 1 we see that the modified Jeffreys procedure is asymptotically minimax and maximin for both formulations of expected regret $E_\theta \log p(x^n|\theta)/q(x^n)$ and $E_\theta \log p(x^n|\hat{\theta})/q(x^n)$ as well as for the pointwise regret.

Remark 6: The constant in the asymptotic minimax regret $C_m = \log((\Gamma(1/2))^m/\Gamma(m/2))$ is also identified in Ordentlich and Cover [24] in a stock market setup and by Freund [15] (for $m = 2$) and Xie [40] (for $m > 2$) using Riemann integration to analyze the Shtarkov value. Szpankowski [33] (see also Kløve [21]) gives expansions of c_n accurate to arbitray order (for $m = 2$). This constant $\log((\Gamma(1/2))^m/\Gamma(m/2))$ can also be determined from examination of an inequality in Shtarkov [30, eq. (15)] and it is given in Shtarkov *et al.* [31]. Here the determination of the constant is a by-product of our principal aim of identifying natural and easily implementable asymptotically maximin and minimax procedures.

Remark 7: Since $\Gamma(1/2) = \sqrt{\pi}$ and

$$\log\Gamma(m/2) = \log(\sqrt{2\pi}(m/2)^{m-1/2}e^{-(m/2)}) + \text{rem}_m$$

by Stirling's approximation to the Gamma function (see [37, p. 253]), an alternative expression for the asymptotic minimax regret from Theorem 1 is

$$r_n = \frac{m-1}{2}\log\frac{n}{m} + \frac{m}{2}\log e - \frac{1}{2}\log 2 - \text{rem}_m + o(1)$$

where $o(1) \to 0$ as $n \to \infty$ and the remainder rem_m in Stirling's approximation is between 0 and $\frac{1}{6m}\log e$. Thus with the remainder terms ignored, the minimax regret equals

$$\frac{m-1}{2}\log\frac{ne}{m}$$

plus a universal constant $\frac{1}{2}\log(e/2)$.

Remark 8: Theorem 2 has implications stochastic lower bounds on regret, that is, lower bounds that hold for most sequences. We use the fact that the Jeffreys' mixture

$$m_J(x^n) = \int w_J(\theta)p(x^n|\theta)\,d\theta$$

using the fixed prior w_J achieves regret never smaller than $\frac{d}{2}\log\frac{n}{2\pi}+C_m$ (which we have shown to be the asymptotically minimax value).

Note that the sequence of mixtures $m_J(x^n)$ is compatible with a distribution M_J on infinite sequences. It follows from [1, Theorem 3.1] (see also [4], [23], and [35] for related conclusions) that for every Kolmogorov-compatible sequence of strategies $q_n(x_1,\cdots,x_n)$ the regret is at least the regret achieved by $m_J(x_1,\cdots,x_n)$ minus a random variable ν depending on $(x_1,x_2,\cdots)$ which for all $t>0$ has the upper probability bound $M_J(\nu>t)\le 2^{-t}$ and, consequently, $P_\theta(\nu>2t)\le 2^{-t}$ for all θ except those in a set of w_J-probability less than 2^{-t}. (In particular, these conclusions hold with $\nu=\sup_n \log q(x^n)/m_J(x^n)$). Thus the regret is never smaller than $\frac{d}{2}\log_2 n+C_m-\nu$, where ν is stochastically dominated by an exponential random variable with mean $2\log_2 e$, for sequences $(x_1,x_2,\cdots)$ distributed according to P_θ for most θ. The implication is that the constant C_m cannot be improved by much in the pointwise sense for most sequences.

Remark 9: As we mentioned in Section I, our priors $\tilde{W}_n$ that provide asymptotically minimax strategies have slight dependence on the sample size n, through the value $\theta^*=b_n/n$ that $J_{i,n}$ sets for the ith cordinate of θ and also through the choice of ε_n. Fortunately, the behavior of the regret is relatively insensitive to the values of b_n and ε_n. In the theorem statement, we set $b_n=1$ and $\varepsilon_n=(1/n)^{1/8}$. A range of choices provide the same conclusions. In particular, the proof will show that if ε_n tends to zero but not too fast, in the sense that $\varepsilon_n\ge n^{-s}$ with $s<1/2$, if b_n is any sequence not greater than $\frac{1}{4}\,(\frac{1}{2}-s)\log n$ and, to prevent b_n from being too small, if $(\log 1/b_n)/\log n\to 0$, then the conclusion of the theorem holds.

An implication is robustness of the procedure to misspecification of the time horizon. Indeed, suppose we set the prior in accordance with an anticipated time horizon N, say with $\theta^*=1/N$ and $\varepsilon=(1/N)^{1/8}$, but for whatever reason compression or prediction stops at time n, with $1/c\le n/N\le c$ for some constant c. Then the resulting procedure still satisfies the conditions for the conclusion of the theorem to hold.

IV. Proof of the Main Theorem

The statements of the theorem and the corollary are based on the following inequalities which we will prove.

$$\frac{m-1}{2}\log\frac{n}{2\pi}+C_m\le\sum_{x^n} m_J(x^n)\log\frac{p(x^n|\hat{\theta})}{m_J(x^n)} \tag{3}$$

$$\begin{aligned}
&\le \max_W \sum_{x^n} m_J(x^n)\log\frac{p(x^n|\hat{\theta})}{m_J(x^n)}\\
&\le \min_q \max_{x^n}\log\frac{p(x^n|\hat{\theta})}{q(x^n)}\\
&\le \max_{x^n}\log\frac{p(x^n|\hat{\theta})}{\tilde{q}(x^n)}\\
&\le \frac{m-1}{2}\log\frac{n}{2\pi}+C_m+o(1)
\end{aligned} \tag{4}$$

where $C_m=\log(\Gamma(1/2)^m/\Gamma(m/2))$. Since both ends in the above are asymptotically equal, it follows that

$$\begin{aligned}
&\frac{m-1}{2}\log\frac{n}{2\pi}+C_m+o(1)\\
&\quad=\sum_{x^n} m_J(x^n)\log\frac{p(x^n|\hat{\theta})}{m_J(x^n)}+o(1)\\
&\quad=\log c_n=\underline{r}_n=\overline{r}_n\\
&\quad=\max_{x^n}\log\frac{p(x^n|\hat{\theta})}{\tilde{q}(x^n)}+o(1)\\
&\quad=\frac{m-1}{2}\log\frac{n}{2\pi}+C_m+o(1)
\end{aligned} \tag{5}$$

and, therefore, $C_m=\log(\Gamma(1/2)^m/\Gamma(m/2))$ is the asymptotic constant in the minimax regret, Jeffreys' mixture m_J is asymptotically maximin (least favorable), and the modified Jeffreys' mixture $\tilde{q}_n$ is asymptotically minimax.

We consider the regret using Jeffreys' mixture $m_J(x^n)$. From Lemma 1 of the Appendix, this regret is asymptotically constant (independent of x^n) for sequences with relative frequency composition internal to the simplex, that is, when $\min(T_1,\cdots,T_m)\to\infty$. However, Lemma 3 exhibits a constant higher regret on vertex points when using Jeffreys' mixture. Thus Jeffreys' mixture is not asymptotically minimax on the whole simplex of relative frequencies.

Now we verify inequalities (3) and (4). The three inequalities between them follow from the definitions and from maximin $\le$ minimax.

The proof for line (3) follows directly from Lemma 2, where it is shown that $\log\frac{p(x^n|\hat{\theta})}{m_J(x^n)}$ is greater than $\frac{m-1}{2}\log\frac{n}{2\pi}+C_m$ for all sequences x^n.

To prove inequality (4) we proceed as follows. We denote the count of symbol i in a sequence x^n by $T_i=T_{i,n}$. Let $\tau_n\ge 1$ be a sequence with $\tau_n\to\infty$. Observe that for x^n in the region of $\mathcal{X}^n$ where $T_i\ge\tau_n$ for all $i=1,\cdots,m$, using the upper bound from Lemma 1 in the Appendix, we have

$$\begin{aligned}
\log\frac{p(x^n|\hat{\theta})}{\tilde{q}_n(x^n)} &< \log\frac{p(x^n|\hat{\theta})}{(1-\varepsilon_n)m_J(x^n)}\\
&\le\left(\log\frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)}+\frac{m-1}{2}\log\frac{n}{2\pi}\right)\\
&\quad+\left(\frac{m}{4\tau_n+2}+\frac{m^2}{4n}\right)\log e+\log\frac{1}{1-\varepsilon_n}\\
&=\log\frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)}+\frac{m-1}{2}\log\frac{n}{2\pi}+o(1)
\end{aligned} \tag{6}$$

where the remainder term in (6) tends to zero uniformly (for sequences with $T_i\ge\tau_n$) as $n\to\infty$. In accordance with Remark 9, let the sequence ε_n be chosen such that $\varepsilon_n\to 0$ with $n^{-s}\le\varepsilon_n\le 1/2$, with exponent $s<1/2$, moreover, let b_n be in the range of values indicated there. Then for our analysis we choose the sequence $\tau_n\to\infty$ in such a way that

$$\tau_n\log\tau_n/b_n\le\tfrac{1}{2}(\tfrac{1}{2}-s)\log n$$

and $\tau_n\ge b_n$. (For instance, if $b_n=1$ a choice of $\tau_n\sim\frac{1}{2}(\frac{1}{2}-s)\log n/\log\log n$ suffices.) Now we consider the region of $\mathcal{X}^n$ where $T_i<\tau_n$ for some i. This region is the union of the subregions where $T_i<\tau_n$ for $i=1,\cdots,m$. For the ith

such subregion we use $\varepsilon_n m_i(x^n)$ to lower-bound $\tilde{q}_n(x^n)$. For notational simplicity take $i = 1$ and denote $b = b_n$. Then

$$m_1(x^n) = \int_{\theta_2+\cdots+\theta_m=1-(b/n)} \left(\frac{b}{n}\right)^{T_1} \cdot \theta_2^{T_2-(1/2)} \cdots \theta_m^{T_m-(1/2)} d\theta_2 \cdots d\theta_{m-1} \Big/ \int_{\theta_2+\cdots+\theta_m=1-(b/n)} \theta_2^{-(1/2)} \cdots \theta_m^{-(1/2)} d\theta_2 \cdots d\theta_{m-1} = \frac{D_{m-1}\left(T_2+\frac{1}{2}, \cdots, T_m+\frac{1}{2}\right) \cdot \left(\frac{b}{n}\right)^{T_1} \left(1-\frac{b}{n}\right)^{n-T_1}}{D_{m-1}\left(\frac{1}{2}, \cdots \frac{1}{2}\right)}.$$

It follows that

$$\log \frac{p(x^n|\hat{\theta})}{\tilde{q}_n(x^n)} \le \log \frac{p(x^n|\hat{\theta})}{\varepsilon_n m_1(x^n)} = \log \frac{1}{\varepsilon_n} + \log \frac{\left(\frac{T_1}{n}\right)^{T_1} \left(\frac{n-T_1}{n}\right)^{n-T_1}}{\left(\frac{b}{n}\right)^{T_1} \left(\frac{n-b}{n}\right)^{n-T_1}} + \log \frac{\prod_{i=2}^{m} \left(\frac{T_i}{n-T_1}\right)^{T_i}}{\frac{D_{m-1}\left(T_2+\frac{1}{2}, \cdots, T_m+\frac{1}{2}\right)}{D_{m-1}\left(\frac{1}{2}, \cdots, \frac{1}{2}\right)}} \quad (7)$$

where if $T_1 = 0$ we follow the convention that $T_1^{T_1} = 1$. When $b \le T_1 \le \tau_n$ this middle term is bounded by $\tau_n \log(\tau_n/b)$, whereas, when $0 \le T_1 \le b$, it is bounded by

$$n \log n/(n-b) = n \log(1 + b/(n-b))$$

which is not more than $2b$. The third term of (7) can be bounded using the conclusion of Lemma 1 of the Appendix for the Jeffreys' mixture (here for the $n - T_1$ observations that do not involve symbol 1). Thus we have

$$\log \frac{p(x^n|\hat{\theta})}{\tilde{q}_n(x^n)} \le \log \frac{1}{\varepsilon_n} + \max\{2b_n, \tau_n \log(\tau_n/b_n)\} + \frac{m-2}{2} \log \frac{n}{2\pi} + \left(\frac{m^2}{4n} + \frac{m}{2}\right) \log e + \log \frac{\left(\Gamma\left(\frac{1}{2}\right)\right)^{m-1}}{\Gamma\left(\frac{m-1}{2}\right)}.$$

By our choices of ε_n, b_n, and τ_n we have that the first term is bounded $s \log n$ and the second term by $\frac{1}{2}\left(\frac{1}{2} - s\right) \log n$, so that the sum of these contributions represents a cost quantifiably less than the $\frac{1}{2} \log n$ saved from the other terms (for sequences with at least one count less than τ_n). We have that for $n \ge m/2$

$$\log \frac{p(x^n|\hat{\theta})}{\tilde{q}_n(x^n)} \le \left(\frac{m-2}{2} + \frac{\frac{1}{2}+s}{2}\right) \log \frac{n}{2\pi} + m \log e + \log \frac{\left(\Gamma\left(\frac{1}{2}\right)\right)^{m-1}}{\Gamma\left(\frac{m-1}{2}\right)} + \frac{1}{2} \log 2\pi. \quad (8)$$

We see that the contribution from the boundary regions produces regret not more than

$$(m - 1/2) \log(n/2\pi) + C_m$$

for all sufficiently large n, in particular for n such that

$$\frac{1}{2}\left(\frac{1}{2} - s\right) \log \frac{n}{2\pi} \ge m \log e + \frac{1}{2} \log 2\pi + \log \frac{\Gamma\left(\frac{m}{2}\right)}{\Gamma\left(\frac{m-1}{2}\right) \Gamma\left(\frac{1}{2}\right)}. \quad (9)$$

Then, putting this near boundary analysis together with the interior analysis from (6), we have that overall our regret exceeds $\frac{m-1}{2} \log \frac{n}{2\pi} + C_m$ by not more than

$$\text{Rem}_n = \left(\frac{m}{4\tau_n} + \frac{m^2}{4n} + 2\varepsilon_n\right) \log e.$$

Typically, this tends to zero at rate $\log \log n / \log n$ in accordance with the behavior of $1/\tau_n$. (A negligibly faster rate of $1/\log n$ is available if τ_n and b_n are set to both be of order $\log n$.)

As we have seen, the asymptotic value is the same for the upper and lower bounds in inequalities (3) through (4). Thus these collapse into asymptotic equalities and the conclusions follow.

Finally, we show that the modification to produce an asymptotically minimax procedure $\tilde{q}_n$ retains the asymptotic least favorable (maximin) property of Jeffreys' mixture. That is,

$$\sum_{x^n} r(\tilde{q}_n, x^n) \tilde{q}_n(x^n) = \log c_n - D(\tilde{q}_n \| q_n^*) = \log c_n + o(1)$$

or, equivalently, $D(\tilde{q}_n \| q_n^*) \to 0$. Indeed, we have

$$D(\tilde{q}_n \| q_n^*) = D((1 - \varepsilon_n) m_J + \varepsilon_n m_{V,n} \| q_n^*)$$

which by convexity is not greater than

$$(1 - \varepsilon_n) D(m_J \| q_n^*) + \varepsilon_n D(m_{V,n} \| q_n^*).$$

We already showed the first term goes to zero. The second term also converges to zero since $D(m_{V,n} \| q_n^*) \le \log c_n$ and $\varepsilon_n \to 0$ faster than logarithmically. Thus $D(\tilde{q}_n \| q_n^*) \to 0$ as $n \to \infty$. □

V. Other Modifications of Jeffreys' Prior

In this section we briefly mention another possible modification of Jeffreys' mixture that can achieve a somewhat faster rate of convergence to the asymptotic minimax value. In Section IV we added some point mass to the Jeffreys' prior near the boundary of the simplex to pull down the regret incurred by sequences with relative frequencies close to or on the boundary. Here instead we add a Dirichlet prior with parameters less that $1/2$.

Consider the modified Jeffreys' prior $W_n^{(2)} = (1 - \varepsilon_n)$ Dirichlet$_m(\frac{1}{2}, \cdots, \frac{1}{2}) + \varepsilon_n$ Dirichlet$_m(\alpha, \cdots, \alpha)$, where $0 < \alpha < 1/2$. As before we let ε_n tend to zero with $\varepsilon_n \ge n^{-s}$. This prior $W_n^{(2)}$ yields a mixture probability mass function $q_n^{(2)}(x^n) = \int p(x^n|\theta) W_n^{(2)} d\theta$ that is also asymptotically minimax. To see why, note first for x^n in the region where $T_i \ge \tau_n$, for all $i = 1, \cdots, m$, we have by the same reasoning as before, that

$$\log \frac{p(x^n|\hat{\theta})}{q_n^{(2)}(x^n)} \le \left(C_m + \frac{m-1}{2} \log \frac{n}{2\pi}\right) + \left(\frac{m^2}{4n} + \frac{m}{4\tau_n} + 2\epsilon_n\right) \log e. \quad (10)$$

Here we set $\tau_n = h^p$ for some $p > 0$. For x^n in the region of $\mathcal{X}^n$ where $T_i < \tau_n = n^p$ for some i, we use Lemma 4 of the Appendix to get that

$$\log \frac{p(x^n|\hat{\theta})}{q_n^{(2)}(x^n)} \le \left(\frac{m-1}{2} - (1/2-\alpha)(1-p)\right)\log n + \left(K_m \log \frac{1}{\alpha} + \log \frac{1}{\varepsilon_n}\right)$$

where K_m is a constant depending only on m. Then as long as $(1/2-\alpha)(1-p) > s$, for large enough n, we have the bound

$$\left(\frac{m-1}{2} - \left(\frac{1}{2}-\alpha\right)(1-p) + s\right)\log n + K_m \log \frac{1}{\alpha} \le \frac{m-1}{2}\log\frac{n}{2\pi} + \log \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)}. \quad (11)$$

Examining (10) and (11), we conclude that for certain choice of p, s, and α, the regret $r(q_n^{(2)}, x^n)$ is asymptotically upper-bounded by $\frac{m-1}{2}\log\frac{n}{2\pi} + C_m + o(1)$, uniformly for all x^n. It is wise to choose $p = s$ (to balance the remainder terms of order $1/\tau_n$ and ε_n in (10)). For example, a choice of $p = s = 1/4$ and $\alpha = 1/8$ satisfies (11). Consequently, $q_n^{(2)}$ is asymptotically minimax. Moreover, the maximal regret converges to the asymptotic minimax value at rate $n^{-1/4}$. A more delicate choice of

$$p_n = s_n = 1/3 - O(\log\log n/(\log n)^2)$$

and $\alpha = K_m/(\log n)^2$ provides for the largest p and s satisfying condition (11) and yields a procedure with maximal regret that converges to the asymptotic minimax value at rate $n^{-1/3}$. These rates may be compared to what is achieved by the exact minimax value $\log c_n$ which for $m = 2$ is shown in [33] to approach the asymptotic value at rate $n^{-1/2}$.

The procedure $q_n^{(2)}(x^n)$ is readily computed using the predictive density

$$q_n^{(2)}(x_{k+1} = j|x^k) = \frac{(1-\varepsilon_n)m_{1/2}(x^{k+1}) + \varepsilon_n m_\alpha(x^{k+1})}{(1-\varepsilon_n)m_{1/2}(x^k) + \varepsilon_n m_\alpha(x^k)}$$

with

$$m_\alpha(x^{k+1}) = m_\alpha(x^k)m_\alpha(x_{k+1}|x^k)$$

where

$$m_\alpha(x_{k+1} = j|x^k) = (T_{k,j} + \alpha)/(k + m\alpha).$$

The total computation time of this iterative algorithm is of order nm.

VI. Application in Gambling

We now study some applications of the main result in this current and the following sections.

Suppose in a horse race we index the horses by $1, \cdots, m$, and we are going to bet on n races. For race k, let the odds be $O_k(x|x_1, \cdots, x_{k-1})$ to 1 for horse x to win. We bet our fortune according to some proportion $q_n(x_k|x_1, \cdots, x_{k-1})$ at game k. Let $X^n = (X_1, \cdots, X_n)$ be the indices of the winning horses. Then the asset at time n would be

$$\begin{aligned} S(X^n, q_n) &= \prod_{k=1}^n (q_n(X_k|X_1, \cdots, X_{k-1}) \\ &\quad \cdot O_k(X_k|X_1, \cdots, X_{k-1})) \\ &= q_n(X_1, \cdots, X_n)O(X_1, \cdots, X_n) \end{aligned}$$

where

$$O(X_1, \cdots, X_n) = \prod_{k=1}^n O_k(X_k|X_1, \cdots, X_{k-1}).$$

If the horse races were random, with outcomes $X_1, \cdots, X_n$, if the win probabilities for each race were $(\theta_1, \cdots, \theta_m)$, and if we knew the parameter θ, we would bet with proportion $q_n(i) = \theta_i$ on horse i (see Cover and Thomas [8, Ch. 6]). Whether or not the races are random, the wealth at time n with such a constant betting strategy θ is

$$\begin{aligned} S(X^n, p_\theta^n) &= \prod_{k=1}^n (p(X_k|\theta)O_k(X_k|X_1, \cdots, X_{k-1})) \\ &= p(X_1, \cdots, X_n|\theta)O(X_1, \cdots, X_n) \end{aligned}$$

where $p(x_1, \cdots, x_n|\theta) = \theta_1^{T_1} \cdot \ldots \cdot \theta_m^{T_m}$ and T_i is the number of wins for horse i. With hindsight, the best of these values is at the maximum likelihood. Hence the ratio of current wealth to the ideal wealth is

$$\begin{aligned} R(X^n, q_n) &= \frac{S(X^n, q_n)}{S(X^n, p_{\hat{\theta}}^n)} \\ &= \frac{q(X_1, \cdots, X_n)O(X_1, \cdots, X_n)}{p(X_1, \cdots, X_n|\hat{\theta})O(X_1, \cdots, X_n)} \\ &= \frac{q_n(X^n)}{p(X^n|\hat{\theta})}. \end{aligned}$$

We want to choose a $q_n(x^n)$ to optimize this ratio, in the worst case. That is, we pick a q_n to achieve

$$\min_{q_n} \max_{\theta, X^n} \log \frac{p(X^n|\theta)}{q_n(X^n)} = \min_{q_n} \max_{X^n} \log \frac{p(X^n|\hat{\theta})}{q_n(X^n)}.$$

This is the quantity our paper has analyzed, and we have provided an asymptotic minimax q_n. We achieve

$$\frac{q_n(X^n)}{p(X^n|\hat{\theta})} \ge C'_m \cdot n^{-(m-1/2)}(1 + o(1)) \quad (12)$$

uniformly for all horse race outcomes X^n, where

$$C'_m = 2^{(m-1)/2}\Gamma(m/2)/\sqrt{\pi}$$

is the best such constant. Here $n^{-(m-1/2)}$ expresses the cost (as a factor of wealth) of the lack of foreknowledge of $\hat{\theta}$. A gambling procedure that achieves (12) is to bet proportion $\tilde{q}(x_{k+1}|x^k)$ of our wealth on the possible outcomes of successive races using the modified Jeffreys' mixture as in (1).

There is an extension of this gambling problem to the stock market with m stocks. In this case

$$S(X^n, q_n) = \prod_{k=1}^n \left(\sum_{i=1}^m q_n(i|X_1, \cdots, X_{k-1})X_{ki}\right)$$

where X_{ki} is the wealth factor (price ratio) for stock i during investment period (day) k and $q(i|x_1, \cdots, x_{k-1})$ is the proportion of wealth invested in stock i at the beginning of day k. Recent work of Cover and Ordentlich [7], [24] shows that for all sequences $x_1, \cdots, x_n$, the minimax log wealth ratio for stocks is the same as the minimax log wealth ratio for horse racing with m horses

$$\min_{q_n} \max_{\theta, x^n} \frac{S(x^n, p_\theta^n)}{S(x^n, q_n)} = \min_{q_n} \max_{x^n} \frac{p(x^n|\hat{\theta})}{q_n(x^n)}$$

where on the left side the maximum is over all $x_1, \cdots, x_n$ with each stock vector x_i in R_+^n and on the right side the maximum is over all $x_1, \cdots, x_n$ with each x_i in $\{1, \cdots, m\}$. Thus from our analysis of the latter problem we have for the stock market that the asymptotic minimax wealth ratio is

$$\min_{q_n} \max_{\theta, x^n} S(x^n, p_\theta^n)/S(x^n, q_n) = n^{(m-1/2)}/C'_m \cdot (1 + o(1))$$

in agreement with Cover and Ordentlich [24]. However, it remains an open problem whether there is an asymptotically minimax strategy that can be evaluated in polynomial time in n and m for the stock market. The best available algorithms in Cover and Ordentlich [24] runs in time of order n^{m-1} compared to time nm^2 obtained here for the horse race case.

VII. Application in Prediction

Suppose we have observed a sequence $x^k = (x_1, \cdots, x_k)$. We want to give a predictive probability function for the next x_{k+1} based on the past i observations, and we denote it by $\hat{p}_k(x|x^k) = q(x|x_1, \cdots, x_k)$ for all $x \in \mathcal{X}$. When x_{k+1} occurs we measure the loss by $\log 1/\hat{p}_k(x_{k+1}|x^k)$. Thus the loss is greater than or equal to zero (and equals zero iff the symbol x_{k+1} that occurs is the one that was predicted with $\hat{p}_k(x_{k+1}|x^k) = 1$). We initiate with a choice $\hat{p}_0(x) = q(x)$ of an arbitrary probability. We denote by

$$q(x_1, \cdots, x_n) = \prod_{k=0}^{n-1} q(x_{k+1}|x_1, \cdots, x_k)$$

the probability mass function obtained as the product of the predictive probabilities. The total cumulative log-loss is

$$\sum_{k=0}^{n-1} \log 1/q(x_{k+1}|x^k) = \log 1/q(x_1, \cdots, x_n). \tag{13}$$

A class

$$p(x_1, \cdots, x_n|\theta) = \prod_{k=1}^{n} p(x_k|\theta), \theta \in \Theta$$

of memoryless predictors incurs cumulative log-loss of

$$\sum_{k=0}^{n-1} \log 1/p(x_k|\theta) = \log 1/p(x_1, \cdots, x_n|\theta)$$

for each θ and with hindsight the best such predictor corresponds to the maximum likelihood. (Using this target class the aim of prediction is not to capture dependence between the $x_1, \cdots, x_n$ but rather to overcome the lack of advance knowledge of $\hat{\theta}$). The log-loss for prediction is chosen for the mathematical convenience of the chain rule (13). Direct evaluation of regret bounds is easier for such a loss than for other loss function. Moreover, log-loss regret provides bounds for minimax regret for certain other natural cumulative loss functions including $0-1$ loss and squared error loss, see [18], [34], and [17]. The minimax cumulative regret is

$$\min_q \max_{\theta, x_1, \cdots, x_n} \sum_{k=0}^{n-1} \log \frac{p(x_{k+1}|\theta)}{q(x_{k+1}|x^k)} = \min_q \max_{x_1, \cdots, x_n} \frac{p(x_1, \cdots, x_n|\hat{\theta})}{q(x_1, \cdots, x_n)}$$

for which we have identified the asymptotics.

The Laplace–Jeffreys update rule is asymptotically maximin and its modification (as given in Theorem 1) is asymptotically minimax for online prediction.

VIII. Application in Data Compression

Shannon's noiseless source coding theory states that for each source distribution $p(x^n|\theta)$, the optimal code length of x^n is $\log 1/p(x^n|\theta)$, ignoring the integer rounding problem (if we do round it up to integer, the extra code length is within one bit of optimum), where in Shannon's theory optimality is defined by minimum expected code length. Kraft's inequality requires that the code-length function $l(x^n)$ of a uniquely decodable code must satisfy $l(x^n) = \log 1/q(x^n)$ for some subprobability $q(x^n)$. When θ is unknown, we use a probability mass function $q(x^n)$ such that for all θ and all x^n, the code length using q is (to the extent possible) close to the smallest of the values $\log 1/p(x^n|\theta)$ over $\theta \in \Theta$. That is, we want to q to achieve

$$\min_q \max_{\theta, x_1, \cdots, x_n} (\log 1/q(x^n) - \log 1/p(x^n|\theta)) = \min_q \max_{x_1, \cdots, x_n} \frac{p(x^n|\hat{\theta})}{q(x^n)}.$$

The choice $q(x^n) = p(x^n|\hat{\theta}(x^n))$ is not available because Kraft's inequality is violated. Shtarkov showed that the minimax optimal choice is the normalized maximum-likelihood

$$q(x^n) = p(x^n|\hat{\theta}) / \sum_{x^n} p(x^n|\hat{\theta}).$$

Implementation of such codes for long block length n would require computation of the marginals and conditionals associated with such a $q(x_1, \cdots, x_n)$. For the normalized maximum likelihood, these conditionals (as required for arithmetic coding) are not easily computed. Instead, we recommend the use of $q(x^n) = m_J(x^n)$ equal to Jeffreys' mixture or its modification, for which the conditionals are more easily calculated (see Remark 3). The arithmetic code for x^n is

$$\overline{F}(x^n) = \sum_{a^n < x^n} q(a^n) + \frac{1}{2} q(x^n)$$

expressed in binary to an accuracy of $\lceil \log(1/q(x^n)) \rceil + 1$ bits. We can recursively update both $F(x^k)$ and $q_n(x^k)$ using the conditionals $q_n(x_k|x_1, \cdots, x_{k-1})$ in the course of the algorithm. For details see [8, pp. 104–107]. We remark here that the distribution constructed in Section V also provides a straightforward algorithm for this arithmetic coding.

IX. Prediction with Side Information

So far we have looked at prediction (as well as gambling and coding) based on a sequence of y_j's with a memoryless target class. In practice, one often has observed some side information x_j to help in the prediction of y_j. In this section, we give the minimax regret for the prediction problem with side information.

Suppose a sequence of data $(x_j, y_j)_{j=1}^n$ is to be observed, where $y_j \in \{1, \cdots, m\}$, and $x_j \in \{1, \cdots, k\}$. We call y_j the response variable and x_j the explanatory variable. We wish to provide a choice of conditional distribution

$$q(y_1, \cdots, y_n | x_1, \cdots, x_n) = \prod_{j=1}^{n} q(y_j | y^{j-1}, x^j)$$

for prediction, gambling, and data compression that perform well compared to a target family of competitors, uniformly over all sequences. The target family of procedures act according to an assumption that $y_1, \cdots, y_n$ are conditionally independent given $x_1, \cdots, x_n$ with the following conditional probability distribution:

$$p(y_j = y | x_j = x) = \theta_{x,y}$$

for $k = 1, \cdots, n$, $y = 1, \cdots, m$ and $x = 1, \cdots, s$. These $\theta_{x,y}$'s are called parameters of the model. Denote the collection of these parameters by θ, that is, $\theta = (\theta_1, \cdots, \theta_k)$ with $\theta_s = (\theta_{s,1}, \cdots, \theta_{s,m})$, for $s = 1, \cdots, m$. (These parameters may be organized into a matrix.) Then the joint conditional probability under the competitor's model can be written as

$$\begin{aligned} p(y_1, \cdots, y_n | x_1, \cdots, x_n, \theta) &= \prod_{j=1}^{n} p(y_j | x_j, \theta) \\ &= \prod_{s=1}^{k} \prod_{j: x_j = s} p(y_j | s, \theta_s) \\ &= \prod_{s=1}^{k} p\left(y^{n_s} | \theta_s\right) \end{aligned}$$

where y^{n_s} is subsequence for which $x_j = s$ (with the understanding that when there are no observations with $x_j = s$, the factor $p(y^{n_s}|\theta_s)$ is set to one so that it has no effect on the product). Here

$$p(y^{n_s} | \theta_s) = \prod_{j: x_j = s} p(y_j | s, \theta_s)$$

treats the observations in this subsequence as if they were independent and identically distributed. The maximum-likelihood estimator is

$$\hat{\theta}_s = \left(\frac{n_{s,1}}{\sum_{i=1}^{m} n_{s,i}}, \cdots, \frac{n_{s,m}}{\sum_{i=1}^{m} n_{s,i}} \right)$$

for $s = 1, \cdots, k$, where

$$n_{s,i} = \sum_{j=1}^{n} 1_{\{x_{j=s}, y_j = i\}}$$

is the number of observations for which the response is i when the explanatory variable is s. We define the regret $r(x^n, y^n, q)$ for using a conditional probability function $q(y^n|x^n)$ as the log ratio between the best of the competitors probability $p(y^n|x^n, \hat{\theta})$ to our choice $q(y^n|x^n)$ at data points (x^n, y^n), that is,

$$r(x^n, y^n, q) = \log \frac{p(y^n | x^n, \hat{\theta})}{q(y^n | x^n)}.$$

We are interested to know the asymptotic minimax value

$$\overline{r}_n = \min_{q(\cdot|\cdot)} \max_{x^n, y^n} r(x^n, y^n, q)$$

and a probability $q(y^n|x^n)$ that asymptotically achieves this minimax value. Moreover, we desire a "causal" q in which the distribution assigned to each y_j depends on past y's and past and present x's but not on future x's.

We will prove that the asymptotic minimax value for the prediction problem with side information is given by

$$\overline{r}_n = \frac{k(m-1)}{2} \log \frac{n/k}{2\pi} + kC_m + o(1).$$

Note that the solution can be interpreted as k times the value we had before, but with n/k in place of n.

The asymptotic upper bound for the minimax value is derived from the following argument. Observe that

$$\overline{r}_n = \min_{q(\cdot|\cdot)} \max_{x^n, y^n} \log \frac{p(y^n | x^n, \hat{\theta})}{q(y^n | x^n)}.$$

For each x^n, let $n_s(x^n) = \{j\colon x_j = s\}$ be the set of indices corresponding to the subsample of observations for which the explanatory variable takes value s (the subsample with context s). With slight abuse of notation we also use n_s to denote the size of this subsample, i.e., the cardinality of $n_s(x^n)$. By choosing q to have the property that

$$q(y^n | x^n) = \prod_{s=1}^{k} q(y^{n_s} | s)$$

where $y^{n_s} = (y_j : j \in n_s)$, we obtain an upper bound on the minimax regret

$$\begin{aligned} \overline{r}_n &\le \max_{x^n} \max_{y^n} \sum_{s=1}^{k} \log \frac{p(y^{n_s} | s, \hat{\theta}_s)}{q(y^{n_s} | s)} \\ &\le \max_{x^n} \sum_{s=1}^{k} \max_{y^n} \log \frac{p(y^{n_s} | s, \hat{\theta}_s)}{q(y^{n_s} | s)}. \end{aligned} \tag{14}$$

The terms in this bound for each subsample is of the type we have studied in this paper. Thus we are motivated to take $q(y^{n_s}|s)$ to be a modified Dirichlet mixture of $p(y^{n_s}|\theta_s)$ for observations in the subsample $n_s(x^n)$. Now the subsample size n_s is not known in advance (though the total sample size n is presumed known). To produce a causal strategy we set values of ε and b (or α) in our modified mixture as if the anticipated subsample sizes were n/k, in a manner that, for realized subsample sizes different from this by not more than a constant factor, tight bounds of the type we have presented still hold. For example, we

may set $\varepsilon = (n/k)^{-1/8}$ and either $\theta^* = k/n$ (for the first modification) or $\alpha = 1/4$ (for the second modification). The regret bound in (14) may be written

$$\overline{r}_n \le \max_{x^n} \left\{ \sum_{\{s:n_s \ge n/kc\}} \max_{y^n} \log \frac{p(y^{n_s}|s,\hat{\theta}_s)}{q(y^{n_s}|s)} + \sum_{\{s:n_s < n/kc\}} \max_{y^n} \log \frac{p(y^{n_s}|s,\hat{\theta}_s)}{q(y^{n_s}|s)} \right\}. \quad (15)$$

Suppose n is large enough that (9) (or its counterpart (11)) is satisfied with $n/(kc)$ in place of n. Then from the result of Section IV (or V) we bound

$$\max_{y^n} \log \frac{p(y^{n_s}|s,\hat{\theta}_s)}{q(y^{n_s}|s)}$$

by $\frac{m-1}{2}\log\frac{n_s}{2\pi} + C_m + \text{rem}_{n/kc}$ for $n_s \ge n/kc$, where the remainder is

$$\text{rem}_{n/kc} = \left(\frac{m^2}{4n/kc} + \frac{m}{4\tau_{n/kc}} + 2\varepsilon\right)\log e$$

which, as we have seen, tends to zero as $n \to 0$. For the cases where $n_s < n/kc$ it is sufficient to use the coarser bound from Lemma 1 of $((m-1)/2)\log(n_s/2\pi) + C_m + (m^2/2)\log e$. Thus we obtain a bound on the regret of

$$\max_{n_1,\cdots,n_k} \left\{ \sum_{s=1}^{k} \left(\frac{m-1}{2}\log\frac{n_s}{2\pi}\right) + kC_m + k\,\text{rem}_{n/(kc)} + \sum_{\{s:n_s<n/kc\}} \frac{m^2}{2}\log e \right\}. \quad (16)$$

The maximum in (16) is over choices of nonnegative $n_1,\cdots,n_k$ that add to n. We shall argue that (with sufficiently large c) the maximum in this bound occurs at $n_s = n/k$. Toward this end we reexpress the bound as

$$\frac{k(m-1)}{2}\log\frac{n/k}{2\pi} + kC_m + k\,\text{rem}_{n/(kc)} \quad (17)$$

$$-\frac{m-1}{2}\min_{q_1,\cdots,q_k}\sum_{s=1}^{k}\left(\log\frac{1/k}{q_s} - 1_{\{q_s<1/kc\}}\frac{m^2}{m-1}\log e\right) \quad (18)$$

where the minimum is over nonnegative $q_1,\cdots,q_k$ that sum to one. Here (17) reveals the desired bound once we show that the minimum in (18) is indeed positive. We recognize the sum in (18) as a multiple of the Kullback divergence between the uniform distribution $1/k,\cdots,1/k$ and $q_1,\cdots,q_k$. Now since these distributions both sum to one, the sum in (17) is unchanged if we add $k(q_s - 1/k)\log e$ to each summand. The new summands are then

$$\left(\log\frac{1}{kq_s} + (kq_s - 1)\log e - 1_{\{kq_s<1/c\}}\frac{m^2}{m-1}\log e\right). \quad (19)$$

We see that this is nonnegative for each s, whether $kq_s \ge 1/c$ (such that the indicator term does not appear) or whether $kq_s < 1/c$, provided c is chosen large enough that $\log c/e \ge (m^2/(m-1))\log e$. The terms in (19) are zero only when $q_s = 1/k$. Thus we have the upper bound on the minimax regret of

$$\frac{k(m-1)}{2}\log\frac{n/k}{2\pi} + kC_m + o(1)$$

as desired.

For a lower bound on $\overline{r}_n$ we use minimax $\ge$ maximin (in fact $\overline{r}_n = \underline{r}_n$ as Theorem 0 shows). The maximin value is

$$\underline{r}_n = \max_{x^n}\max_{p(y^n|x^n)}\min_{q(\cdot|x^n)}\sum_{y^n} p(y^n|x^n)\log\frac{p(y^n|x^n,\hat{\theta})}{q(y^n|x^n)} = \max_{x^n}\max_{p(y^n|x^n)}\sum_{y^n} p(y^n|x^n)\log\frac{p(y^n|x^n,\hat{\theta})}{p(y^n|x^n)}. \quad (20)$$

We obtain a lower bound in (20) by choosing for each x^n

$$p^*(y^n|x^n) = \prod_{s=1}^{k} p^*(y^{n_s}|s)$$

where $p^*(y^{n_s}|s)$ is the mixture of $p(y^{n_s}|\theta_s)$ with respect to the Dirichlet $(1/2,\cdots,1/2)$ prior. Then from Lemma 2 of the Appendix, we have that

$$\log\frac{p(y^n|x^n,\hat{\theta})}{p^*(y^n|x^n)} = \sum_{s=1}^{k}\log\frac{p(y^{n_s}|s,\hat{\theta}_s)}{p^*_{n_s}(y^{n_s})} \ge \sum_{s=1}^{k}\left(\frac{m-1}{2}\log\frac{n_s}{2\pi} + C_m\right).$$

Hence continuing from (20), we have

$$\underline{r}_n \ge \max_{x^n}\sum_{s=1}^{k}\left(\frac{m-1}{2}\log\frac{n_s}{2\pi} + C_m\right) = \frac{k(m-1)}{2}\log\frac{n}{2k\pi} + kC_m.$$

Thus we have shown that the asymptotic minimax regret is

$$r_n = \frac{k(m-1)}{2}\log\frac{n}{2k\pi} + kC_m + o(1).$$

Note that in the upper bound we found a causal $q(y^n|x^n)$ that is asymptotically minimax. By causality we mean that q satisfies

$$q(y^n|x^n) = \prod_{j=1}^{n} q(y_j|x^j, y^{j-1}).$$

Here it is not necessary to condition on future x values as in the general decomposition

$$q(y^n|x^n) = \prod_{j=1}^{n} q(y_j|x^n, y^{j-1}).$$

Moreover, the conditional distribution of y_j given x^j and y^{j-1} depends only on the subsample of past y_i of which $x_i = s$ when $x_j = s$. The advantage of using such a q is that we can give an "online" prediction as data are revealed to us.

APPENDIX

Lemma 1: (Uniform bound for the log-ratio of maximum likelihood and Jeffreys' mixture). Suppose $p(x^n|\theta_1,\cdots,\theta_m) =$

$\theta_1^{T_1} \cdot \ldots \cdot \theta_m^{T_m}$, where T_i is the count for the ith symbol in the alphabet, and $m_J(x^n)$ is Jeffreys' mixture, that is,

$$m_J(x^n) = \int_S p(x^n|\theta_1, \cdots, \theta_m) \cdot \theta_1^{-1/2} \cdot \ldots \cdot \theta_m^{-1/2} \, d\theta_1 \cdots \, d\theta_{m-1}$$

where

$$S = \left\{ (\theta_1, \cdots, \theta_{m-1}) : \theta_i \geq 0, \sum_{i=1}^{m-1} \theta_i \leq 1 \right\}.$$

Then for all x^n, we have

$$\log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} = \frac{m-1}{2} \log \frac{n}{2\pi} + C_m + Res_n \qquad (21)$$

where

$$C_m = \log \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)}$$

and

$$0 < Res_n \leq \left(\frac{m^2}{4n} + \frac{m}{4 \min(T_1, \cdots, T_m) + 2} \right) \log e. \qquad (22)$$

In particular

$$\log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} \leq \frac{m-1}{2} \log \frac{n}{2\pi} + C_m + \left(\frac{m^2}{4n} + \frac{m}{2} \right) \log e \qquad (23)$$

which, to state a somewhat cruder bound, is not greater than $(m-1/2)\log(n/2\pi) + C_m + (m^2/2)\log n$, valid for all $m \geq 2$ and $n \geq 1$.

Note: Expression (22) shows that we have an accurate characterization of regret in the interior of the relative frequency simplex. On the full simplex, the bound in (23) is somewhat larger (as it must be since the regret at each vertex of the relative frequency simplex, corresponding to a constant sequence, is higher than in the interior, see Lemma 3). Similar bounds for Jeffreys' mixture in the $m = 2$ case are in Freund [15]. We use inequality (23) with a modification of Jeffreys' prior on a reduced dimension simplex in the proof of the main theorem.

Proof: We leave the lower bound proof to Lemma 2 and only prove the upper bound here. By Stirling's formula for real-valued $x > 0$ (see [37, p. 253])

$$\Gamma(x) = x^{x-1/2} e^{-x} \sqrt{2\pi} e^s \qquad (24)$$

where the remainder $s = s(x)$ satisfies $0 < s < 1/(12x)$. Thus Jeffreys' mixture $m_J(x^n)$ can be approximated as the following:

$$\begin{aligned} m_J(x^n) &= D_m\left(T_1 + \frac{1}{2} \cdots, T_m + \frac{1}{2}\right) \Big/ D_m\left(\frac{1}{2}, \cdots, \frac{1}{2}\right) \\ &= \frac{\prod_{i=1}^m \Gamma\left(T_i + \frac{1}{2}\right)}{\Gamma\left(n + \frac{m}{2}\right)} \Big/ \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} \\ &= \frac{\prod_{i=1}^m \left(\sqrt{2\pi}\left(T_i + \frac{1}{2}\right)^{T_i}\right)}{\sqrt{2\pi}\left(n + \frac{m}{2}\right)^{n+(m-1/2)}} \frac{\prod_{i=1}^m e^{s_i}}{e^{s_n}} \Big/ \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} \end{aligned}$$

where the remainders $s_i = s(T_i + 1/2)$ and $s_n = s(n + 1/2)$ are bounded by $1/(12T_i + 6)$ and $1/(12n + 6)$, respectively. Hence

$$\begin{aligned} \log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} &= \log \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} + \frac{m-1}{2} \log \frac{n}{2\pi} \\ &\quad + \left(s_n - \sum_{i=1}^m s_i \right) \log e + \frac{m-1}{2} \ln\left(1 + \frac{m}{2n}\right) \\ &\quad + n \ln\left(1 + \frac{m}{2n}\right) - \sum_{i=1}^m T_i \ln\left(1 + \frac{1}{2T_i}\right) \qquad (25) \end{aligned}$$

where, collectively, the remainder term from the Stirling's approximation satisfies

$$s_n - \sum_{i=1}^m s_i < \frac{1}{12n+6}. \qquad (26)$$

Other remainder terms in (25) are analyzed in the following. We use the following inequality, valid for positive x:

$$\frac{1}{2} - \frac{1}{4(x+1/2)} \leq x \ln\left(1 + \frac{1}{2x}\right) \leq \frac{1}{2} \qquad (27)$$

to get that

$$\begin{aligned} &\frac{m-1}{2} \ln\left(1 + \frac{m}{2n}\right) + n \ln\left(1 + \frac{m}{2n}\right) \\ &\quad - \sum_{i=1}^m T_i \ln\left(1 + \frac{1}{2T_i}\right) \\ &\leq \frac{m(m-1)}{4n} + \frac{m}{2} - \sum_{i=1}^m T_{\min} \ln\left(1 + \frac{1}{2T_{\min}}\right) \\ &\leq \frac{m^2}{4n} + \frac{m}{4T_{\min} + 2} \qquad (28) \end{aligned}$$

where $T_{\min} = \min(T_1, \cdots, T_m)$. Summation of (26) and (28) yields the upper bound in (22). Thus continuing from (25) we obtain that

$$\log \frac{p(x^n|\hat{\theta})}{m_J(x^n)} = \log \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} + \frac{m-1}{2} \log \frac{n}{2\pi} + \mathrm{Res}_n$$

with Res_n satisfying the upper bound in (22) (the lower bound $\mathrm{Res}_n > 0$ is shown in Lemma 2). Inequality (23) follows using $T_{\min} \geq 0$. □

Lemma 2: (A uniform lower bound for the log-ratio of maximum likelihood and Jeffreys's mixture). Using the same notation as in Lemma 1, we have $\mathrm{Res}_n > 0$. Moreover,

$$\log p(x^n|\hat{\theta})/m_J(x^n) - (m - 1/2)\log(n/2\pi)$$

is a decreasing function of the counts $T_1, \cdots, T_m$.

Proof: Define

$$f(T_1, \cdots, T_m) = \frac{p(x^n|\hat{\theta})}{m_J(x^n) n^{(m-1)/2}}$$

where $n = \sum_{i=1}^m T_i$. Once we show that f is decreasing in each variable, it will then follow that

$$\begin{aligned} f(T_1, \cdots, T_m) &> f(T_{\max}, \cdots, T_{\max}) \\ &\geq \lim_{L\to\infty} f(L, \cdots, L) \\ &= \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} \Big/ (2\pi)^{m-1/2} \end{aligned} \tag{29}$$

where $T_{\max} = \max(T_1, \cdots, T_m)$, from which it follows that $R_n > 0$. Now we show that

$$f(T_1 + 1, T_2, \cdots, T_m) < f(T_1, T_2, \cdots, T_m).$$

We have

$$\begin{aligned} &f(T_1, T_2, \cdots, T_m) \\ &\quad = \frac{\left\{\Gamma\left(\frac{1}{2}\right)^m / \Gamma(m/2)\right\} \cdot \left(\prod_{i=1}^m T_i^{T_i}\right) / n^n}{\left\{\left(\prod_{i=1}^m \Gamma\left(T_i + \frac{1}{2}\right)\right) / \Gamma\left(n + \frac{m}{2}\right)\right\} \cdot n^{m-1/2}} \\ &\quad = f(T_1 + 1, T_2 \cdots, T_m) \frac{\left(T_1 + \frac{1}{2}\right) T_1^{T_1}}{(1 + T_1)^{1+T_1}} \\ &\qquad \cdot \frac{(n+1)^{n+1+(m-1/2)}}{\left(n + \frac{m}{2}\right) \cdot n^{n+(m-1/2)}}. \end{aligned} \tag{30}$$

The factor $(T_1 + \frac{1}{2})T_1^{T_1}/(1 + T_1)^{1+T_1}$ is decreasing in T_1 as seen by examining its logarithm. Indeed,

$$g(t) = \log(t + \tfrac{1}{2}) + t \log t - (t+1)\log(t+1)$$

has derivative

$$g'(t) = (t + \tfrac{1}{2})^{-1} + \log(t/(t+1))$$

which (upon setting $t + \frac{1}{2} = (1/2u)$) equals $2u + \log((1 - u)/(1 + u))$, which is negative by examination of the Taylor expansion of $\log(1 + u)$. Consequently, replacing T_1 with n in this factor, we obtain

$$\begin{aligned} &\frac{\left(T_1 + \frac{1}{2}\right) T_1^{T_1}}{(1+T_1)^{1+T_1}} \frac{(n+1)^{n+1+(m-1/2)}}{\left(n + \frac{m}{2}\right) \cdot n^{n+(m-1/2)}} \\ &\quad \geq \frac{\left(n + \frac{1}{2}\right) n^n}{(1+n)^{1+n}} \frac{(n+1)^{n+1+(m-1/2)}}{\left(n + \frac{m}{2}\right) \cdot n^{n+(m-1/2)}} \\ &\quad = \frac{n + \frac{1}{2}}{n + \frac{m}{2}} \left(1 + \frac{1}{n}\right)^{m-1/2} \\ &\quad > 1, \end{aligned} \tag{31}$$

where (31) is equivalent to

$$(n + \tfrac{1}{2})^2 (1 + (1/n))^{m-1} > (n + (m/2))^2$$

which is verified using the binomial expansion of $(1 + (1/n))^{m-1}$. Recalling (30), we have shown that

$$f(T_1, T_2, \cdots, T_m) > f(T_1 + 1, T_2, \cdots, T_m)$$

so it is decreasing in T_1. The same arguments show that f is decreasing in each of the counts.

Finally, the limit of $f(L, \cdots, L)$ as $L \to \infty$ is obtained from

$$f(L, \cdots, L) = \frac{(1/m)^{mL}}{\left\{\Gamma\left(L + \frac{1}{2}\right) / \Gamma\left(mL + \frac{m}{2}\right)\right\}\left\{\Gamma\left(\frac{m}{2}\right) / \Gamma\left(\frac{1}{2}\right)^m\right\}(mL)^{m-1/2}}$$

and then using Stirling's approximation. □

Note: A similar monotonicity argument is given in [38] for the $m = 2$ case.

Lemma 3: (Asymptotic regret on vertex points). At the vertices of the frequency composition simplex (such as $T_1 = n$, and $T_i = 0$ for $i = 2, \cdots, m$), the regret of the Jeffreys' mixture is higher than the asymptotic regret in the interior.

Proof: On the vertex $(n, 0, \cdots, 0)$ we have

$$\begin{aligned} \log \frac{p(x^n|\hat{\theta}_1)}{m_J(x^n)} &= \log \frac{1}{D_k\left(n + \frac{1}{2}, \frac{1}{2}, \cdots, \frac{1}{2}\right) / D_k\left(\frac{1}{2}, \cdots, \frac{1}{2}\right)} \\ &= \log \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} - \log \frac{\Gamma\left(n + \frac{1}{2}\right)\Gamma\left(\frac{1}{2}\right)^{m-1}}{\Gamma\left(n + \frac{m}{2}\right)} \\ &= \log \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} + \frac{m-1}{2} \log \frac{n}{\pi} + o(1) \end{aligned}$$

see also Suzuki [32] and Freund [15]. The asymptotic regret for interior point is

$$\log \frac{\Gamma\left(\frac{1}{2}\right)^m}{\Gamma\left(\frac{m}{2}\right)} + \frac{m-1}{2} \log \frac{n}{2\pi} + o(1)$$

(in agreement with $r_n = \log c_n$). Thus the regret on the vertex is larger by the amount $\frac{m-1}{2} \log 2$, asymptotically. □

Lemma 4: (Regret incurred by other Dirichlet mixtures). Suppose that $\alpha < 1/2$ and let

$$m_\alpha(x^n) = D_m(T_1 + \alpha, \cdots, T_m + \alpha)/D_m(\alpha, \cdots, \alpha).$$

Suppose $n \geq n$. If $T_i < n^p$ for some $i \leq m$ and some $p < 1$, then

$$\log \frac{p(x^n|\hat{\theta})}{m_\alpha(x^n)} \leq \left(\frac{m-1}{2} - \left(\frac{1}{2} - \alpha\right)(1-p)\right) \log n + K_m \log \frac{1}{\alpha}$$

where K_m is a constant depending only on m.

Proof: Without loss of generality we assume that $T_1 < n^p$. Stirling's formula gives the following expansion:

$$m_\alpha(x^n) = \frac{\prod_{i=1}^m (\sqrt{2\pi}(T_i + \alpha)^{T_i + \alpha - 1/2})}{\sqrt{2\pi}(n + m\alpha)^{n + m\alpha - 1/2} \cdot D_m(\alpha, \cdots, \alpha)} \cdot e^R$$

where

$$R = \sum_{i=1}^m s(T_i + \alpha) - s(n + m\alpha)$$

is the residual from the Stirling approximation and thus satisfies

$$\begin{aligned} R &\geq -\frac{1}{12(n + m\alpha)} \\ &\geq -\frac{1}{12n}. \end{aligned} \tag{32}$$

Take the logarithm to get

$$\begin{aligned}\log \frac{p(x^n|\hat{\theta})}{m_\alpha(x^n)} \le & -\frac{m-1}{2}\log(2\pi) \\ & -\sum T_i \log\left(1+\frac{\alpha}{T_i}\right) \\ & +\left(\frac{1}{2}-\alpha\right)\sum_{i=1}^{m}\log(T_i+\alpha) \\ & +n\log\left(1+\frac{m\alpha}{n}\right)-R\log e \\ & +\left(m\alpha-\frac{1}{2}\right)\log(n+m\alpha) \\ & +\log D_m(\alpha,\cdots,\alpha). \end{aligned} \tag{33}$$

To further bound (33) we use

$$\begin{aligned}\sum \log(T_i+\alpha) &= \log(T_1+\alpha) \\ &\quad +\sum_{i=2}^{m}\log(T_i+\alpha) \\ &\le \log(T_1+\alpha)+(m-1)\log\left(\frac{n-T_1}{m-1}+\alpha\right) \\ &\le p\log n+\alpha+(m-1)\log n+\frac{(m-1)^2}{2n}.\end{aligned}$$

Meanwhile, we use $\sum T_i \log(1+\alpha/T_i) > 0$ and $\log(1+x) \le x$ to simplify some other terms in (33). Collectively these yield an upper bound for $\log p(x^n|\hat{\theta})/m_\alpha(x^n)$

$$\log \frac{p(x^n|\hat{\theta})}{m_\alpha(x^n)} \le \left(\frac{m-1}{2}-\left(\frac{1}{2}-\alpha\right)(1-p)\right)\log n + b \tag{34}$$

where the constant b satisfies

$$b \le \left(\frac{(m-1)^2}{4n}+\frac{1}{12n}+\frac{m(m+1)}{4}\right)\log e+\log D_m(\alpha,\cdots,\alpha).$$

By Stirling's approximation,

$$\begin{aligned}D_m(\alpha,\cdots,\alpha) &= \frac{\Gamma(\alpha)^m}{\Gamma(m\alpha)} \\ &\le (2\pi)^{(m-1)/2}\alpha^{1/2-m/2}m^{-m\alpha+1/2}\end{aligned}$$

hence there exists some K_m such that

$$b \le K_m \log\frac{1}{\alpha}.$$

This completes the proof. □

ACKNOWLEDGMENT

The authors wish to thank T. Cover, E. Ordentlich, Y. Freund, M. Feder, Y. Shtarkov, N. Merhav, and I. Csiszár for helpful discussions regarding this work.

REFERENCES

[1] A. R. Barron, "Logically smooth density estimation," Ph.D. dissertation, Stanford Univ., Stanford, CA, 1985.
[2] ——, "Are Bayes rules consistent in information," in *Open Problems in Communication and Computing*, T. M. Cover and B. Gopinath, Eds., 1987, pp. 85–91.
[3] ——, "Exponential convergence of posterior probabilities with implications for Bayes estimations of density functions," Dept. Stat., Univ. Illinois, Urbana–Champaign, IL, Tech. Rep. 7, 1987.
[4] A. R. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.
[5] B. S. Clarke and A. R. Barron, "Information-theoreetic asymptotics of Bayes methods," *IEEE Trans. Inform. Theory*, vol. 36, pp. 453–471, May 1990.
[6] ——, "Jeffreys' prior is asymptotically least favorable under entropy risk," *J. Statist. Planning Inference*, vol. 41, pp. 37–60, Aug. 1994.
[7] T. M. Cover and E. Ordentlich, "Universal portfolios with side information," *IEEE Trans. Inform. Theory*, vol. 42, pp. 348–363, Mar. 1996.
[8] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[9] L. D. Davisson and A. Leon-Garcia, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, Mar. 1980.
[10] ——, "A source matching approach to finding minimax codes," *IEEE Trans. Inform. Theory*, vol. IT-26, pp. 166–174, Mar. 1980.
[11] L. D. Davisson, R. J. McEliece, M. B. Pursley, and M. S. Wallace, "Efficient universal noiseless source codes," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 269–279, May 1981.
[12] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1268, July 1992.
[13] W. Feller, *An Introduction to Probability Theory and Its Applications*. New York: Wiley , 1968, vol. I, 3rd ed..
[14] D. P. Foster, "Prediction in the worst case," *Ann. Statist.* , vol. 19, no. 2, pp. 1084–1090, 1991.
[15] Y. Freund, "Predicting a binary sequence almost as well as the optimal biased coin," in *Proc. 9th Annu. Workshop Computational Learning Theory*, 1996, pp. 89–98.
[16] D. Haussler, "A general minimax result for relative entropy," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1276–1280, 1997.
[17] D. Haussler and A. R. Barron, "How well do Bayes methods work for on-line prediction of $\{\pm 1\}$ values?," in *Proc. NEC Symp. Computation and Cognition*, 1992.
[18] D. Haussler, J. Kivinen, and M. K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1906–1925, Sept. 1998.
[19] D. Haussler and M. Opper, "Mutual information, metric entropy and cumulative relative entropy risk," *Ann. Statist.*, vol. 25, pp. 2451–2492, Dec. 1997.
[20] H. Jeffreys, *Theory of Probability*. Oxford, U.K.: Oxford Univ. Press, 1961.
[21] T. Kløve, "Bounds on the worst case probability of undetected error," *IEEE Trans. Inform. Theory*, vol. 41, pp. 298–300, Jan. 1995.
[22] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 199–207, Jan., 1981.
[23] N. Merhav and M. Feder, "Universal Sequential Decision Schemes from Individual Sequences," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 1280–1292, July 1993.
[24] E. Ordentlich and T. M. Cover, "The cost of achieving the best portfolio in hindsight," *Math. Oper. Res.*, 1998.
[25] M. Opper and D. Haussler, "Worst case prediction over sequences under log loss," in *The Mathematics of Information Coding, Extraction, and Distribution*. New York: Springer Verlag, 1997.
[26] J. Rissanen, "Universal coding, information, prediction and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, 1984.
[27] ——, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526–532, July 1986.
[28] ——, "Fisher information and stochastic complexity," *IEEE Trans. Inform. Theory*, vol. 40, pp. 40–47, Jan 1996.
[29] Yu. M. Shtarkov, "Encoding of Discrete Sources Under Conditions of Real Restrictions and Data Compression," Speciality 05.13.01, Technical cybernetics and information theory, Ph.D. dissertation, Moscow, USSR, 1980.
[30] ——, "Universal sequential coding of single messages," *Probl. Inform. Transm.* , vol. 23, pp. 3–17, July 1988.
[31] Yu. M. Shtarkov, Tj. J. Tjalkens, and F. M. J. Willems, "Multi-alphabet university coding of memoryless sources," *Probl. Inform. Transm.* , vol. 31, pp. 114–127, 1995.
[32] J. Suzuki, "Some notes on universal noiseless coding," *IEICE Trans. Fundamentals*, vol. E78-A, , no. 12, Dec. 1995.
[33] W. Szpankowski, "On asymptotics of certain sums arising in coding theory," *IEEE Trans. Inform. Theory*, vol. 41, no. 6, pp. 2087–2090, Nov. 1995.

[34] V. Vovk, "Aggregating strategies," in *Proc. 3rd Annu. Workshop Computer Learning Theory*, 1990, pp. 371–383.

[35] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, pp. 384–396, Mar. 1994.

[36] M. J. Weinberger, N. Rissanen, and M. Feder, "A universal finite memory sources," *IEEE Trans. Inform. Theory*, vol. 41, pp. 643–652, May 1995.

[37] E. T. Whittaker and G. N. Watson, *A Course of Modern Analysis*, 4 ed. Cambridge, U.K.: Cambridge Univ. Press, 1963.

[38] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "The context tree weighting method: Basic properties," *IEEE Trans. Inform. Theory*, vol. 41, pp. 653–664, May 1995.

[39] Q. Xie and A. R. Barron, "Minimax redundancy for the class of memoryless sources," *IEEE Trans. Inform. Theory*, vol. 43, pp. 646–657, May 1997.

[40] Q. Xie, "Minimax Coding and Prediction," Ph.D. dissertation, Yale Univ., New Haven, CT, May 1997.